Construction d'un système lexical multilingue, libre de droits, centré sur le français et le japonais via des méthodes automatiques et contributives

Mathieu Mangeot

Mathieu MANGEOT-NAGATA

Université Hosei, 3-7-2 Kajinocho, Koganei-shi, TOKYO 184-8584 Japon
Université de Savoie, Laboratoire d'Informatique de Grenoble BP 53 F-38041 GRENOBLE CEDEX 9 France

Construction d'un système lexical multilingue, libre de droits, centré sur le français et le japonais via des méthodes automatiques et contributives

Contexte

Ce projet se situe dans le domaine du traitement automatique des langues (TAL), à la croisée de l'informatique et de la linguistique, plus précisément sur la lexicographie et la lexicologie multilingues.
Lors d'un premier long séjour au Japon de novembre 2001 à mars 2004, nous avions fait le constat que les ressources lexicales français-japonais disponibles sur le Web étaient quasi inexistantes. Ce qui avait donné naissance au projet Papillon de construction d'une base lexicale multilingue à structure pivot (Mangeot et al., 2003). Depuis, des progrès ont été faits dans plusieurs domaines (technique, théorique, social) (Mangeot, 2006) mais la production concrète de données a peu progressé. D'autre part, la réutilisation de ressources lexicales est à la mode (désambiguïsation lexicale, utilisation de ressources en source ouverte (Wiktionary, dbpedia), fusion avec des ontologies, etc.). Même si elles permettent de consolider et d'élargir la couverture des ressources existantes, ces expériences partent toujours de données créées à la main par des lexicographes.

Problématique

Bien que le français et le japonais soient considérées comme des langues bien dotées au niveau des outils et des ressources linguistiques, le couple français-japonais est considéré comme un couple de langues peu doté. Il existe en effet peu de ressources lexicales bilingues électroniques de qualité et libres de droits. Les corpus bilingues alignés et les systèmes de traduction automatique français-japonais sont logiquement tout aussi rares.
Pour des raisons historiques autant que pratiques, les Japonais ont mis rapidement l'accent sur l'anglais. Le couple anglais-japonais est donc l'un des mieux dotés à l'heure actuelle avec des ressources très conséquentes comme le dictionnaire EDR (1993) et des systèmes de traduction automatique parmi les plus performants.
Les dictionnaires japonais-français existants de bonne qualité sont des dictionnaires éditoriaux qui n'existent qu'au format papier ou compilé dans des dictionnaires électroniques (denshi-jishou). Il n'existe pas d'interface de consultation en ligne.
Pour le français->japonais, il existe par exemple Le Dico (Hakusuisha, 1993) avec 34 000 entrées et le Crown (Sanseido) avec 47 000 entrées. Pour le japonais->français, il existe par exemple le Royal (Obunsha, 1992) avec 42 000 entrées et le Concise (Sanseido) avec 38 000 entrées.
Il existe un dictionnaire français-japonais disponible en ligne, il s'agit du projet dictionnaire-japonais.com[1] (ou Nichifutsu Jiten) qui contient un peu plus de 28 000 mots. Il constitue un net progrès par rapport aux autres projets de dictionnaire japonais-français en ligne. Chaque utilisateur peut contribuer directement en rajoutant des entrées. Les informations disponibles pour chaque entrée sont relativement limitées à un "type grammatical", une "catégorie" (domaine), un registre de langue, et parfois une "origine du mot" (étymologie).
Il contient par contre très peu d'exemples et n'est semble-t'il pas disponible au téléchargement.
Pour l'anglais et l'allemand, il existe par contre des dictionnaires bilingues de bonne couverture et de bonne qualité disponibles en ligne et surtout en téléchargement. Il s'agit pour l'anglais du projet JMdict dirigé par Jim Breen qui contient actuellement et environ 160 000 entrées pour l'allemand du projet WaDokuJiten dirigé par Ulrich Apel qui contient environ 280 000 entrées.

Matériel et Méthode

Partant de ce constat, nous avons défini le projet suivant qui consiste à construire un système lexical multilingue riche d'informations avec priorité sur le couple de langues français-japonais.
Ce système lexical sera constitué d'une part d'un corpus bilingue aligné français-japonais et d'autre part d'un dictionnaire bilingue (dans un premier temps) à structure pivot.
Le corpus bilingue sera constitué de textes enrichis avec des outils d'analyse automatique (lemmes et catégories grammaticales). Le premier objectif du corpus est de servir de source pour trouver des exemples qui enrichiront les entrées du dictionnaire. Il peut également être utilisé pour d'autres buts : construction d'un système de traduction automatique statistique, lexicométrie, étude de textes, etc.
La construction du dictionnaire se fera d'une part par la réutilisation de ressources existantes (lexiques franco-japonais, Wiktionary) et leur exploitation automatique (réification de liens de traduction, désambiguïsation de sens de mots) et d'autre part par des contributeurs bénévoles travaillant en communauté sur le Web. Ceux-ci seront amenés à contribuer soit via des jeux lexicaux sérieux, soit directement sur les articles de dictionnaire en fonction de leur niveau d'expertise et de leurs connaissances dans le domaine de la lexicographie ou de la traduction bilingue.
La microstructure des articles composant les volumes monolingues est une simplification de celle du projet Papillon. Chaque article est cette fois basé sur le vocable. Un vocable étant soit un regroupement de lexies (sens de mot), soit une locution.

Les lexies sont constituées d'un nom, des propriétés grammaticales, d'une formule sémantique qui peut être vue comme une définition formelle - dans le cas d'une lexie, prédicative, la formule décrit le prédicat et ses arguments et on trouve

1 http://dictionnaire-japonais.com

aussi le régime qui décrit la réalisation syntaxique des arguments - , puis d'une liste de fonctions lexico-sémantiques - il y a 56 fonctions de base applicable à toute langue et pouvant se combiner entre elles -, d'une liste d'exemples et enfin d'une liste d'expressions idiomatiques.
Pour faire face aux niveaux de compétences différents selon les contributeurs, l'interface d'édition pourra s'adapter et afficher une granularité d'information adaptée. Par exemple, un contributeur débutant sera invité à renseigner une simple glose pour caractériser une lexie, alors qu'un linguiste expert devra décrire une formule sémantique complète. De même, certains contributeurs seulement auront accès à la liste des fonctions lexicales à remplir.

La macrostructure est également tirée du projet Papillon avec un volume monolingue pour chaque langue et un volume pivot au centre. Cette macrostructure a été expérimentée et validée dans le projet LexAlp1 (Sérasset et al., 2006) de construction d'une terminologie multilingue pour le vocabulaire de la convention alpine. Ce projet utilise également la plate-forme Jibiki (Mangeot et al., 2004) comme pour son développement et sa consultation en ligne.
Lorsqu'un nouvel article dans une langue A est ajouté, il doit être relié au volume interlingue. Ces liens sont créés soit en réutilisant des dictionnaires bilingues existants langue A→langue B, soit en les ajoutant manuellement à partir d'une traduction. Le lien langue A→langue B devient langue A→pivot→langue B. Si l'article langue B est déjà relié à un autre article langue C, alors l'article langue A bénéficiera lui aussi de ces liens.
Cependant, afin de ne pas dérouter les utilisateurs, ceux-ci contribueront via une interface présentant une vue classique de dictionnaire bilingue. Chaque lien bilingue langue A→langue B ajouté via cette interface sera en fait traduit en arrière plan par la création de deux liens interlingues ainsi que d'une axie représentant le lien de traduction d'origine pour obtenir finalement : langue A→axie pivot→langue B. Cette idée a été utilisée pour le projet MotÀMot (Mangeot, 2014).

Chaque partie d'information de chaque article se verra attribuer un niveau de qualité . Les niveaux s'échelonnent de 1 étoile pour un brouillon (données récupérées dont la qualité n'est pas connue) à 5 étoiles, qualité certifiée par un expert (par exemple, un lien de traduction validé par un traducteur assermenté).
De la même manière, les contributeurs se verront assigner un niveau de compétence (1 à 5 étoiles également). 1 étoile étant le niveau d'un débutant inconnu dans la communauté et 5 étoiles étant le niveau d'un expert reconnu.

Les ressources ainsi produites seront libres de droits, disponibles en téléchargement public et destinées à être utilisées aussi bien par des humains via des dictionnaires bilingues classiques que par des machines pour des outils de traitement automatique de la langue (analyse, traduction automatique, etc.).

Concernant le corpus, l'outil IMS Corpus WorkBench sera utilisé pour l'indexation et le consultation. Les textes français seront analysés par l'analyseur de Brill (TreeTagger) ainsi que par l'étiqueteur morphosyntaxique Melt. Les textes japonais seront analysés par l'outil MeCab.

Concernant le dictionnaire, la plate-forme Jibiki, fondée sur Enhydra, un serveur d'objets Java et Postgresql, déjà utilisée avec succès pour plusieurs autres projet de dictionnaires (GDEF, MotàMot, DiLAF[2]) constituera la base du système lexical.

Résultats/ Discussion

Pour le corpus, un premier site Web expérimental a été lancé[3]. Il regroupe déjà quelques corpus disponibles dont la plupart sont tirés du projet OPUS[4] : une convention de non double-imposition (17 000 mots), la bible (900 000 mots), le coran (192 000 mots), OpenOffice (250 000 mots) et OpenSubtitles (4 millions de mots). Nous cherchons actuellement des textes bilingues pour étoffer notre collection. Pour enrichir le corpus, il est important de collecter tous les genres de textes (littérature, textes légaux, Web, etc.). Si vous avez connaissance d'une telle ressource, n'hésitez pas à nous contacter.
Pour le dictionnaire, nous sommes également dans une phase de collection de lexiques bilingues existants. De même, si vous avez connaissance d'une telle ressource, n'hésitez pas à nous contacter. D'ici peu, nous fusionnerons ces lexiques pour constituer un squelette de dictionnaire qui sera ensuite à réviser et compléter.
Les aspects techniques du projet ont été longuement abordés et sont pratiquement maîtrisés. Nous sommes otpimistes quant à la réalisation d'un premier brouillon de dictionnaire. Par contre, le succès réel d'un tel projet réside plutôt dans l'étape d'après qui implique l'adhésion d'une communauté de contributeurs bénévoles. C'est pourquoi nous voulons déjà proposer des ressources utilisables avant d'ouvrir le site au public.

Liste de publications

Mangeot Mathieu, Sérasset Gilles & Lafourcade Mathieu **Construction collaborative d'une base lexicale multilingue.** Traitement Automatique des Langues, vol. 44(2), pp. 151–176, 2003.

Mangeot Mathieu **Papillon project : Retrospective and perspectives.** In P. Zweigenbaum, Ed., Acquiring and Representing Multilingual, Specialized Lexicons : the Case of Biomedicine, LREC workshop, Genoa, Italy, 6 p, 2006.

Mathieu Mangeot **MotàMot project conversion of a French-Khmer published dictionary for building a multilingual lexical system.** Languages Resources and Evaluation Conference, May 2014, Reykjavik, Iceland. pp.8, 2014.

2 http://www.estfra.ee, http://jibiki.univ-savoie.fr/motamot, http://www.dilaf.org

3 http://jibiki.univ-savoie.fr/uplug/corpus.pl

4 http://opus.lingfil.uu.se